09/029831

METHOD FOR GENERATION OF AN N-WORD PHRASE DICTIONARY FROM A TEXT CORPUS

ABSTRACT

It is, therefore, an object of the present invention to provide a structure and method for automatically creating a dictionary for clustering text documents, including performing a first pass for each of the documents to determine a frequency of each word in each of the documents, creating a Hashtable of most frequently occuring words in the documents, performing a second pass for each of the documents to determine a frequency of phrases in each of the documents that contain only words in the Hashtable and adding the most frequently occuring phrases to the Hashtable, and outputting the most frequently occuring words and the most frequently occuring phrases as the dictionary. The determination of the frequency of each word can include removing punctuation and case from the documents, removing stop words from the document, replacing words in the documents with synonyms, removing duplicate words from the documents, adding remaining words to the Hashtable, determining the frequency of each word remaining in the Hashtable, and removing words below a frequency level from the Hashtable.